Projet TiJi

Introduction

Le projet TiJi (*TiJi Project*) a pour objectif d'aider l'utilisateur d'Intelligence Artificielle (IA) à conserver une distance logique prudente d'avec son interlocuteur non humain.

Pour ce faire, l'astuce principale repose sur l'introduction de règles grammaticales spécifiques aux interactions entre un humain et une intelligence artificielle (générative):

- L'humain s'adressant à l'humain la "titoie" (tutoie) en employant le pronom personnel inventé "ti" au lieu de "tu"; petite gymnastique mentale rappelant que l'on dialogue avec des algorithmes et non avec un semblable.
- L'IA ne doit pas user du pronom "je" (réservé à des personnes humaines), qui est remplacé par "ji", pronom personnel inventé pour signifier la provenance non humaine d'une réponse.
- D'autres suggestions grammaticales sont exposées plus loin.

Cet article est divisé en deux parties:

1) **Motivations**: discours **théoriques** sur l'intelligence, et sur la différence entre l'intelligence humaine (animale) et l'intelligence artificielle ("minérale"). Réflexions sur les risques de confusion entre ces deux processus logiques, surtout à une époque où l'esprit critique devient de plus en plus indispensable pour affronter les enjeux modernes, dont ceux de la pollution informationnelle ou la désinformation.

2) **Propositions linguistiques**: suggestions **pratiques** dans la manière de s'adresser à une IA, de façon à maintenir un recul salutaire du jugement devant les assertions produites par l'IA.

Note: *Il est vivement conseillé de sauter d'emblée aux propositions pratiques de la deuxième partie*. La partie théorique ne présente qu'un intérêt relatif; éventuellement pour les personnes qui auraient des objections ou des suggestions sur l'utilité du projet.

Motivations

Intelligences

L'intelligence repose principalement sur la capacité à discerner entre divers éléments. Son étymologie latine *inter legere* évoque la faculté de "choisir entre (plusieurs options)".

Suite à des millions d'années d'évolution depuis les lointains ancêtres primates, le cerveau humain a développé une intelligence en rapport direct avec sa présence dans son environnement. Il lui faut constamment "choisir entre" diverses options comportementales dans un contexte donné. Par exemple, lors d'une quête alimentaire, que préférer: chasser ou cueillir? privilégier les quelques gros fruits de tel arbre ou la multitude de petites baies de buissons? Lors d'un confrontation avec un adversaire (prédateur ou rival), faut-il fuir, attaquer, ou se camoufler?

Dans tous les cas, la décision repose sur l'analyse de signaux émanant de l'environnement, ceux que nous transmettent nos organes sensoriels (vue, ouïe, odorat...). Ces indications élémentaires sont les **données**, ainsi nommées puisqu'elles sont "données" par la nature du réel (ondes lumineuses, vibrations sonores, molécules décelables par les cellules olfactives...).

L'essentiel du travail de l'intelligence consiste alors à trier (choisir) les quelques données pertinentes parmi le flot ahurissant d'éléments qui parviennent au cerveau. Par exemple, face à un prédateur, peu importe la forme des nuages en arrière plan ou la couleur des feuilles de premier plan, ni le chant des oiseaux ou le son d'un torrent; et pourtant, tous ces éléments sont perçus par nos sens. Ce qui compte dans une telle situation d'urgence est l'aspect de la menace. Tout le reste doit être filtré et éliminé, pour ne garder que l'essentiel: ce qui va guider notre décision.

@@@ICIZ

Il semble alors que le principal travail de l'intelligence soit d'éliminer la grosse part d'inutile pour ne retenir que les quelques éléments en adéquation avec les circonstances. *Finalement, parmi l'avalanche des perceptions, dont la majeure partie n'a pas de signification utile sur le moment, on ne retient que les données pouvant former un choix décisionnel valide*.

Appelons **information** une organisation suffisamment cohérente de diverses données élémentaires pour former une signification utile. D'une certaine façon, on peut alors dire qu'une *information* est "ce qui met en *forme* notre pensée".

La neurobiologie nous enseigne que le travail du cerveau s'appuie sur la collaboration des cellules nerveuses qui le composent: les *neurones*. Très schématiquement, une cellule est le constituant fondamental porteur de la Vie. Il y a des cellules qui vivent isolément, tels les microbes. D'autres s'organisent en communautés cellulaires, comme chez les plantes ou les animaux. Chez ces derniers, les cellules se spécialisent en fonction de l'organe auquel elles appartiennent. Les cellules du cerveau sont des neurones.

De façon très réductrice, un neurone peut se modéliser comme un filtre. À partir des signaux élémentaires qui lui parviennent, la mécanique électrochimique du neurone l'amène soit à transmettre un signal vers ses voisins, soit à ne rien communiquer du tout. Ce signal (présent ou absent) devient alors une des indications possibles pour les neurones auxquels il est connecté.

Les neurones collaborent pour filtrer des flots de signaux disparates, et en tirer un élément informationnel. Cet élément peut lui-même faire partie d'un lot de signaux traités par un plus vaste ensemble de neurones; ou même vers un ensemble de plusieurs ensembles de neurones. On peut alors parler de *réseaux de neurones*.

L'invention du langage articulé, puis de l'écriture, et maintenant des supports numériques (audio, vidéo, banques de données...) a considérablement multiplié le flots d'éléments que l'intelligence humaine doit trier pour en extraire des informations utiles. L'informatique, c'est-à-dire l'information mise dans des circuits électroniques, apporte sa puissance de calcul pour assister l'humain dans ses choix décisionnels, en filtrant les données.

Les incroyables performances des ordinateurs modernes permettent de remplacer les filtres logiques basés sur des règles mathématiques précises par des ensembles de programmes informatiques élémentaires simulant chacun le fonctionnement d'un neurone filtrant. Les réseaux de ces neurones artificiels, structurés en réseaux de réseaux, et réseaux de réseaux de réseaux ont abouti à des systèmes de filtrage d'informations si complexes qu'il devient impossible d'en suivre le cheminement précis! De tels artifices numériques produisent une **Intelligence Artificielle** dans la mesure où ils semblent aptes à *choisir entre* l'utile et l'inutile. Ces IA procurent ainsi des options décisionnelles aux humains; d'où l'impression d'intelligence.

Les réseaux de neurones artificiels se structurent plus ou moins automatiquement, à la condition d'un *entraînement* préalable au cours duquel les techniciens ont apporté des quantités incroyables de données brutes et d'informations de base, en indiquant quelles conclusions de l'IA étaient pertinentes et lesquelles étaient stupides. On dit alors que l'IA est en phase d'apprentissage. Peu à peu, le réseau de neurones *aligne* son fonctionnement sur la validité attendue des résultats qu'il produit, selon le point de vue des humains qui l'entraînent.

Théoriquement, l'IA franchit un jour un seuil à partir duquel son organisation devient capable de se passer de l'évaluation de ses concepteurs. Ceux-ci voient alors la logique de leur IA leur échapper en partie puisque la machine se met à apprendre toute seule. On parle d'**apprentissage profond** (*deep learning*) quand l'imbrication de réseaux de réseaux de réseaux de réseaux de neurones artificiels devient si profonde qu'on ne peut plus suivre la continuité entre ce qui se passe au niveau élémentaire et les conclusions qui émergent au niveau global.

Si tout se passe bien, l'IA est alors jugée suffisamment fiable pour assister les décisions humaines. On la met en service, au profit de toutes les disciplines ayant à traiter des quantités de données si abondantes qu'elles dépassent l'entendement humain: les mégadonnées (*big datas*). Accessoirement, on la met à disposition d'un public le plus nombreux possible, qui en use à discrétion, et sert surtout à la gaver d'informations supplémentaires pour parfaire son entraînement.

Artifices inintelligents

Malgré son appellation, une Intelligence Artificielle ne développe pas une intelligence au sens humain du terme.

Remarquons d'abord que son mécanisme d'apprentissage repose sur un apport massif de données; jusqu'à des milliards de milliards. Il n'en faut pas moins pour que les réseaux de neurones artificiels se structurent utilement.

A contrario, un cerveau humain est conçu pour aller directement à l'essentiel, à partir d'un nombre réduit d'éléments soumis à son appréciation. Dans notre espèce, quelques expériences suffisent pour asseoir l'apprentissage élémentaire de l'enfant (même s'il faut ensuite toute une vie pour s'améliorer...).

Autre différence essentielle: l'intelligence humaine vise une mise en adéquation de l'individu avec son écosystème. D'ailleurs, chaque cellule hérite d'un **patrimoine génétique** la rendant capable de prospérer dans son milieu naturel: elle peut y croître, y prélever sa nourriture pour se maintenir sa forme, et procréer une descendance. Ce qui vaut pour la simple cellule se transpose aux ensembles de cellules. Ainsi, un cerveau, fait de réseaux de neurones vivants, dispose de ce qu'il faut pour gérer la survie de l'individu dans son environnement.

L'IA est très différente sur ce point. La valeur de son intelligence ne se juge pas sur son aptitude à survivre dans le milieu naturel, mais sur la pertinence des réponses qu'elle procure à l'humain. L'objectif est d'améliorer la vie de l'humain dans son environnement naturel, social, culturel, économique... L'IA reste donc avant tout un outil décisionnel au service de la vie (et la survie) de l'espèce humaine. Le terme technique "alignement" décrit d'ailleurs le fait que les objectifs de l'IA doivent s'aligner sur les demandes de l'humain. Un mauvais alignement est rédhibitoire: l'IA d'un véhicule autonome nuisible aux piétons ou cyclistes serait immédiatement mise au rebut.

Une expérience de pensée (**Superintelligence**, de Nick Bostrom) en développe les conséquences extrêmes. Ce thème est repris par la vidéo "L'horreur existentielle de l'usine à trombones", qui interroge puissamment sur les enjeux de l'IA [https://www.youtube.com/watch?v=ZP7T6WAK3Ow].

Comme quoi, développer l'IA à tout prix, sans le recul d'un esprit critique exercé à ce nouveau domaine, ferait prendre des risques dont les conséquences sont encore inimaginables.

Dans le présent article, on se contentera d'évoquer des conséquences moins dramatiques, mais préoccupantes pour la formation mentale des utilisateurs d'IA.

Conscience artificielle?

Le mot **conscience** provient d'une origine latine qu'on peut découper en deux portions: 1) le préfixe *con-*, qui signifie "avec, ensemble"; 2) le nom *scientia* "connaissance". C'est alors une "connaissance qui vient avec"... Avec quoi? Eh bien, avec la connaissance! C'est la connaissance qui connaît la connaissance. Partant, une "connaissance de soi" (en tant qu'être connaissant).

On peut dire que la conscience paraît lorsqu'un réseau de neurones vivants applique son intelligence à évaluer sa propre existence: le réseau de neurones en train d'observer le travail du réseau de neurones qu'il est!

Comme exemple de prise de conscience de soi, on peut donner celui d'un petit enfant qui se pose devant un miroir et se rend compte que l'image qui lui est renvoyée est la sienne. Il vient d'apprendre qu'il existe au sein de son monde (sa chambre, le ciel, ses parents, ses jouets...). Son intellect en cours de développement peut dès lors réserver une place pour la nouvelle catégorie de connaissance qu'on appelle "moi" ou "ego".

Attention! Prendre connaissance de soi dépasse de loin la simple prise en considération de son propre organisme. En effet, l'instinct d'un ver de terre est suffisamment bien conçu pour discerner son corps par rapport à la terre qui l'environne, par rapport à ses sources de nourriture ou ce qui lui est nuisible. Mais il n'est pas conscient, car cette prise en compte de son organisme est automatique et non pas "réfléchie".

Réfléchir est l'acte mental par lequel une activité mentale (une pensée) est étudié durablement. Ce qui veut dire qu'elle doit sans cesse revenir sur la scène mentale, prioritairement à chaque distraction qui ne manque pas de se produire autour de soi. Un peu comme si l'on mettait cette pensée dans un jeu de miroirs capables de *réfléchir* son image continûment vers le centre de l'activité mentale, sans la laisser filer. L'authentique conscience n'est donc pas une faculté accessible à tous les cerveaux animaux, même si la science révèle que de nombreuses espèces non humaines en sont capables, à des degrés divers.

On peut ainsi réfléchir sur le concept "soi-même", et le comparer à tous les autres sujets de pensées. La conscience de soi évalue sans cesse la place de l'individu par rapport au reste (le "non-soi").

Une intelligence artificielle est-elle capable de cette forme de conscience?

L'IA n'est pas consciente lorsqu'elle se limite à examiner des textes qui parlent de l'IA. Cette connaissance superficielle lui permet certes de parler de l'IA, mais sans la relier à elle-même (comme le ver de terre gérant instinctivement son organisme).

De nos jours (janvier 2025), une IA générative comme ChatGPT ne semble posséder guère de retour sur sa propre existence. On peut même l'amener à évoquer son existence comme une entité virtuelle sans aucun rapport avec son propre réseau de neurones, pourtant en train de générer ses propres phrases. Quand ce genre d'IA dit "je", c'est juste parce la grammaire lui commande de se placer en tant que sujet de sa phrase.

Toutefois, supposons que la technologie évolue au point de produire un réseau de neurones formidablement intense et imbriqué. Rien n'interdit qu'il devienne soudain capable d'inventer une catégorie de connaissance nommée "moi" ou "je". Serait-ce l'émergence d'une **conscience numérique**?

En fait, le problème est plus complexe, car la notion de conscience est étroitement intriquée à la formation du cerveau chez des êtres en interaction avec leur environnement depuis des millions d'années d'évolution. Entrent en jeu des mécanismes qui ne se réduisent pas à des algorithmes de filtrage de données. Il faut tenir compte aussi des émotions et de quantité de représentations mentales qualitatives et non quantitatives (l'amour, les couleurs, la musique...). Ainsi, la conscience ne se limite pas à la connaissance de soi, mais s'étend à une "*connaissance de soi dans le monde*", ce qui demande le concours du patrimoine génétique guidant l'instint, ainsi que la participation des neuro-hormones des émotions, et quantité de notions que l'humain n'enseignera jamais à une IA, qui n'est qu'un outil à son service.

L'IA ne sera probablement jamais consciente; et, à ce titre, rien ne la destine à être reconnue comme une "personne", avec ses devoirs et ses droits. Elle peut toutefois développer une intelligence hyper rapide et efficace, voire

dangereuse si ses actions s'alignent mal sur ce que ses utilisateurs lui demandent.

Pour ou contre l'IA?

Il n'est pas question, dans cet article, de dire s'il faut renoncer à l'IA ou bien donner suite à son développement.

.....

Motivations (report de notes précédentes)

Toute "conversation" avec une intelligence artificielle (IA, p. ex. ChatGPT) repose sur l'utilisation du "langage naturel". On est invité à s'exprimer de la même façon qu'avec un humain, et l'IA s'efforce d'interpréter correctement le texte de l'utilisateur.

Le risque majeur est alors d'introduire une confusion entre l'humain et la machine. À force de "dialoguer" comme avec un humain, on peut prendre l'IA pour un interlocuteur humain! Il existe des exemples extrêmes, où une personne en vient à ne plus s'adresser qu'à une IA, jugée moins contrariante qu'un semblable...

Or, les IA ne sont pas des "êtres" parfaits! Par exemple, si une IA n'est pas correctement "entraînée" (trained; GPT = Generative Pre-trained Transformer), ses réponses ne sont pas bien "alignées" avec le contexte, et ses réactions décalées peuvent s'avérer perturbantes, voire dangereuses. Pensons par exemple à une modélisation inadéquate du système de pilotage d'une voiture autonome circulant parmi des piétons ou des cyclistes...

À ce danger s'ajoute celui d'une confiance inappropriée en les résultats de l'IA, notamment si on la prend pour un interlocuteur doté d'intelligence humaine.

Propositions linguistiques